UNIVERSITAT POLITÈCNICA DE CATALUNYA

FINAL DEGREE THESIS

Fake News Classificator

Author:
Elena Ruiz Cano

Director:
Javier Béjar

January 24, 2019

Abstract

Nowadays fake news are considered a problem for the world of information. The objective of this project is to research about this type of news and its main characteristics in order to be able to detect them automatically. This research will focus on classifying false news according to the style and the content. Finally, a web service will be implemented where will include one of the implemented classifiers in order to make predictions about the content of online articles and, at the same time, to retrain the classifier with the articles that it could not predict correctly.

Abstract

Hoy en día las noticias falsas son consideradas un problema para el mundo de la información. El objetivo de este proyecto es investigar en que consisten este tipo de noticias y sus características principales para poder detectarlas de manera automática. Para ello, esta investigación se centrará en clasificar las noticias falsas según el estilo y el contenido. Finalmente se implementará un servicio web que incluya uno de los clasificadores implementados para realizar predicciones de artículos de contenido en linea y a la vez re entrenarse con los artículos que no ha podido predecir correctamente.

Abstract

Avui en dia, les notícies falses es consideren un problema dins del món de la informació. L'objectiu d'aquest projecte és investigar en què consisteixen aquest tipus de notícies i les seves característiques principals per poder detectar-les de manera automàtica. Per això, aquesta investigació se centrarà a classificar les notícies en funció del seu estil i contingut. Finalment s'implementarà un servei web, que inclourà un dels classificadors implementats, per realitzar prediccions d'articles en línia, i alhora reentrenar-se amb articles que el sistema no ha pogut predir correctament.

Contents

Abstract **2**

1 Introduction **6**

2 State of art: Fake News **7**
2.1 Definition . . . 7
2.2 Other type of articles . . . 7
2.3 Objectives . . . 8
2.4 Characteristics . . . 8
2.4.1 Style and Content . . . 9
2.5 How to combat fake news . . . 9
2.5.1 Social Media companies . . . 9
2.5.2 Fact-check organisations . . . 10

3 Project scope **11**
3.1 Motivation . . . 11
3.2 Objectives . . . 11
3.3 Project process . . . 12
3.4 Adquired knowledges . . . 13
3.4.1 Scrapping . . . 13
3.4.2 Natural language process . . . 13
3.4.3 Binary classifiers . . . 13
3.4.4 Use and creation of web services . . . 14

4 Methodology **15**
4.1 Chosen methodology . . . 15
4.2 Risks . . . 16
4.3 Alterations . . . 16

5 Design and implementation **18**
5.1 Architecture . . . 18
5.2 Theoretical methods . . . 19
5.2.1 Binary classification with SVM . . . 19
5.2.2 Dimensional reductions . . . 20
5.2.3 Natural Language Processing . . . 21
5.2.4 Preprocessing . . . 21
5.2.5 Vector transformation with TF-IDF . . . 23
5.2.6 Topic modelling with Latent Dirichlet Allocation . . . 24
5.3 Tools . . . 24

5.3.1 Scikit-learn and Genism 24
5.3.2 Jupyter notebook 25

6 Dataset 26
6.1 Dataset selection 26
6.1.1 Option 1: Search for existing datasets 26
6.1.2 Option 2: Generate a dataset 28
6.1.3 Arised problems 29
6.1.4 Conclusions 29
6.2 Selected dataset 30
6.2.1 List of articles 30
6.2.2 Process of content collection 30
6.2.3 Exploration 31

7 Analysis and classification based on the style of the articles 32
7.1 Objectives of the experiment 32
7.2 Implementation 32
7.2.1 Data extraction 32
7.2.2 Data exploration 34
7.2.3 Training and validation datasets 36
7.2.4 Direct classification with Support Vector Machine 36
7.2.5 Reduced dimensions with Principal Component Analysis and Linear Discriminant Analysis 37
7.2.6 Classification with reduced dimension data 38

8 Analysis and classification based on the content of the articles 40
8.1 Experiment objectives 40
8.2 Implementation 40
8.2.1 Data extraction 40
8.2.2 Data exploration 41
8.2.3 Training and validation datasets 42
8.2.4 Classification with TF-IDF and cosine similarity 42
8.2.5 Classification from Latent Dirichlet Allocation topic distribution 47

9 Web Service 50
9.1 Introduction 50
9.2 Design 50
9.2.1 Objetives 50
9.2.2 Architecture 51
9.2.3 Classifier 52
9.3 Implementation 52

9.3.1 Folder structure . . . 52
9.3.2 Methods . . . 53
9.4 Conclusions . . . 54

10 Project planning **55**
10.1 Schedule . . . 55
10.1.1 Calendar . . . 55
10.1.2 Tasks . . . 55
10.1.3 GANTT Diagram . . . 57
10.2 Alternatives and action plan . . . 58
10.2.1 Learning process . . . 58
10.2.2 Instability in the effort of hours . . . 59
10.3 Changes from the initial planning . . . 59
10.3.1 Delay on scheduling . . . 59
10.3.2 Change in the serialization of some tasks . . . 59
10.3.3 Final schedule . . . 60

11 Budget **62**
11.1 Budget grouping . . . 62
11.1.1 Hardware budget . . . 62
11.1.2 Software budget . . . 62
11.1.3 Human resources budget . . . 63
11.1.4 Unexpected costs . . . 64
11.1.5 Other general costs . . . 64
11.2 Total budget . . . 65

12 Sostenibility **65**
12.1 Enviormental dimension . . . 66
12.2 Economic dimension . . . 66
12.3 Social dimension . . . 66

13 Conclusions **68**
13.1 Acquired knowledge . . . 68
13.2 Project results . . . 68

14 Future work **70**

1 Introduction

This project is a Final Degree Project for the Degree in Computer Engineering of the Faculty of Computer Science of Barcelona. The purpose of this project is to carry out a study on *fake news* and to be able to implement a system that can classify them.

Fake news is taken a key role in the current model information. In front of a globalised model, where people can be easily informed, a lot of people have seen a loudspeaker on social networks in the way of disinforming.

Fake News can have different purposes, but all of them have in common that they want to drive people to read those news as much as possible. Besides, their origin are not fortuitous, a lot of people use this type of news as a business and they end up discrediting the journalistic model.

The concept of fake news has always existed for hundreds of years, but until now no action has been taken so far. The reason is that actually the impact is much bigger than before because currently people can decide the information that want to consume. Moreover, if people don't work to combat them, in the future will be more news of this kind.

This project will attempt to address in depth the main differences between fake and real news to be detected automatically and contribute to a small solution to this major problem.

2 State of art: Fake News

Fake news are being successful because they are often difficult to differentiate from the real ones. In this section, we will try to understand a bit more the reasons.

2.1 Definition

Today there is no consent on the definition of fake news, a fact that generates more confusion when talking about them. For that reason, it is important to talk about the disinformation first, where includes the *malinformation and misinformation.* [2]

The word disinformation has two different interpretations, (I) Giving manipulated information deliberately to serve specific purposes; (II) Giving insufficient information or omitting it.

So there are two sides of the disinformation. On the one hand, the action of malinforming deliberately of non-existent information. On the other hand, the fact of biasing information misinforming.

This fact could imply to the fake news definition which its main objective is disinforming. In conclusion, there are two variants of fake news: (I) Any news that provides false information, even knowing it is not real. (II) Any news that omits important information or biases the context to give a different idea of what happened.

2.2 Other type of articles

In order to know more about the fake news, it is important to understand what is not considered. On this section other type of news that disinform by with other purposes will be defined, and also its differences with fake news.

1. **Propaganda**
   Propaganda is created to convince. This kind of news is based on valid information with the difference that the article is subjective.[4] So when the article is biased, it can omit some information or modify the context but not enough for considering it as fake news.
2. **Satirical**
   Satirical news provides false information deliberately, with the big difference

that the objective is to entertain the reader[5]. Moreover, both, the reader and the writer know that this news are not authentic.

2.3 Objectives

Another point to focus on fake news is to know the different objectives that they have. Taking into account that its definition has different readings, with its objectives happens the same. The objectives below are grouped, as defined in Fake News: The truth of fake news [6], into three groups: economic, ideological and entertainment purposes.

1. **Economic purpose**
   Fake news with an economic purpose wants to become viral and then earn money thanks to the visits. Creating an impacting or controversial news could give more visibility on the article.
2. **Ideological purpose**
   Ideological news is where the author provides subjective information in order to try to convince about their ideology. With this purpose, they include provenly false facts. It is in that moment when the news goes from propaganda to fake news.
3. **Entertainment purpose**
   This news have the objective of entertaining when the person that writes them wants to see the reactions or to observe how the article becomes viral, just for fun.

2.4 Characteristics

As has been said, fake news is a problem that arose suddenly even though it has existed for many years. It is still difficult to have a clear idea about what characteristics the fake news have, specially since they are evolving at the same time as they are spreading through social networks.

Thereupon this section explains some characteristics that fake news can present. These characteristics are grouped, as the *Fake News: A Survey of Research, Detection Methods, and Opportunities*[19] paper poses, by style and content, publication and repercussion. But then the part that will most affect the project, style and content, will be discussed in more depth.

2.4.1 Style and Content

Style refers to the way of expressing a series of ideas and the content is those ideas. These aspects, which depend only on the author, are fundamental in the fake news creation. Both are perfectly accurate in order to reach the main objective: To influence the reader for talking about these articles and to reach the maximum impact possible.

Regarding the style of an article, in most fake news, on the first view, it is possible to see striking titles[12]. This technique is used in order to know what the text is about or to only send a particular message to the readers. Also, the typical formality of a newspaper article descends with the goal of maintaining closer contact with the reader.

About the content aspect, they tend to deal with sensitive problems[13]. These problems usually are timeless to becoming viral, but always the main purpose is to discuss a topic that disturbs the readers.

Another characteristic to add, about the whole style and content aspects, is the bias used. This bias wants to raise concrete ideas for convincing or affecting the reader.

Therefore, when it is trying to detect fake news from the whole article, can be analysed both in a stylistic way as the facts reported.

2.5 How to combat fake news

The problem of fake news is an issue that has reached the current European political scene where even a discussion was opened for figuring out on how to deal with them[8]. As a consequence of its influence, it has been proposed the fact of how we are working in order to prevent its creation and propagation.

2.5.1 Social Media companies

One of the areas which has been most affected by this problem is the social networking companies. Fake news on social media can impact people in a few seconds, and this reason is why fake news has become a serious problem. This fact endangers the credibility of companies like Facebook, Google or Twitter where, apart from seeing the misuse of their networks, they observe how their users

are moving from being connected and informed to ending up uninformed or even misinformed. This risk may become after a while a disinterest in those networks.

In front of this situation, social media companies have started to take actions. Facebook is one of the examples which has opened many offices around the world just for moderating the content possibly false. Youtube, from Google, is another case that is working on avoiding to show content that could be fake. In order to achieve this automatically, they made a previous study on how the media is and its propagation.[11]

Even so, these are the first steps of a long story. It is still difficult to detect them automatically and with the same speed that they propagate where they can pass again unnoticed by the network.

2.5.2 Fact-check organisations

An interesting aspect, since the increasing impact of fake news, is the creation of institutions specially designed to battle this type of articles. These institutions are grouped by the International Fact-Checking Network (IFCN)[17] which includes a code of principles to combat the disinformation.

These organisations are constantly working to refute or affirm, after a rigorous analysis, certain viral news or the ones that concern people. The IFNC gathers organisations from all over the world usually specialise in issues from their origin country, but they can also work on specific subjects such as PoliticFact[16] which deals with news of the U.S. policy.

In Spain two organisations as Maldita.es[1] and Newtral.com[3] can be found. Apart from informing about fake news, in both cases, they have a communication channel, open for everyone, where can personally check those articles which there are doubts about them and then, they affirm or deny with proven facts this information as a free service.

3 Project scope

This section will detail all the objectives of the project and also the processes for achieving them.

3.1 Motivation

Currently, there is a global problem related to fake news propagation, which appears the disinformation as a consequence. Only the people who have the intention to be informed with rigorous criteria can usually detect this type of articles easier. Even so, it is hard to recognise them.

This is why the project will focus on a study that identifies the differences between fake and real news. This study will perform different experiments with the purpose to find a way to recognise, with a good exactitude, fake news from the real ones. Besides, the best classification founded will be included in a system in order to predict the veracity of new articles and to be able to learn from them.

3.2 Objectives

Starting from the motivation, the following objectives have been defined for the project:

1. **Research possible differences between true and fake news**
   Firstly, a set of articles, previously tagged, will be collected for processing and interpreting them in order to extract useful information about them. Depends on the information that will be gotten the experiments will be focused in different ways.
2. **Try to classify the article using different methods**
   From a set of processed data, previously chosen in a justified way, to different transformation techniques and binary classification models using methods for classifying them with the best results.

   This research will be separated in two points of view: classifying using the article style and using its content. So different types of classification techniques will be applied depending on the case.

3. **Generate a self-learning system**
   Implement a software system that allows the users to check an orientation about the reliability of a given determined journalistic article. So thanks to the use of this application the model could improve its training and then getting better results.

3.3 Project process

Regarding the established objectives, the following tasks have been defined in order to achieve them correctly.

1. **Investigate the main characteristics of fake news.**
   Research about what defines fake news and how they can distinguish from real, to be able to get interesting information from the set of articles.
2. **Collect a group of classified articles by fake or real**
   Obtain a set of articles previously tagged as true or false. This process can be done in two different ways, obtaining datasets from third parties or generating our dataset. With these two options, evaluating which is the best option during the project realisation.
3. **Explore and classify the articles from style part**
   Exploring, from the set of articles, the way of extracting qualitative information about documents styles with the help of natural language pre-processing methods.

   Given this extracted information apply different techniques of modulation and classification. Finally, compare the different implemented methods and evaluate which has had better results.
4. **Explore and classify the articles from content part**
   Apply different natural language processes, further that the pre-processing techniques, where text vectorization and topic modelling are included.

   Classify the obtained data, from the mentioned transformations, with supervised classification models to finally collect the training and validation results.
5. **Compare the different classification approaches and assess the results**
   Evaluate, from all used methods, the different obtained results. Then conclude deciding which technique has given a better effect in solving the project problem.

6. **Generate a web application to validate new articles with the chosen model.**
   From the chosen model on the previous task, develop a web application which allows the model to be trained for every request and, at the same time, provide to the user an orientation about the truth of the consulted article.

3.4 Adquired knowledges

All the different topics about computing science and software development that will be applied in this project will be explained below.

3.4.1 Scrapping

One of the project objectives is to obtain a dataset. So in order to collect the data for this dataset the web scrapping method will be used. This method allows to extract information from certain websites. In our case, the title, subtitle and body from each online article will be collected.

3.4.2 Natural language process

Natural Language Processing includes multiples methods to process text from data. These methods will be used to understand and transform it to finally introduce it into classification models.

Some examples of those methods are pre-processing methods that try to standardise the data for searching better results. But also there are processes for creating models representations and topic modelling to extract other types of information.

3.4.3 Binary classifiers

Supervised algorithms for binary classification should be used when performing some classification methods. With these algorithms, models will be created in order to train and validate with the available data.

During this project 'Support Vector Machine' (SVM) model will be used. It is based on Suppor Vector classification idea, where its objective is to draw borders in the space for grouping the data by classes.

3.4.4 Use and creation of web services

Different web services of third parties will be used and will help with the implementation of the project.

In addition, a web service will be implemented in order to apply a model classification in a real case. This web service will consist in returning the truth, of a consulted article, using the best model implemented during the experiments.

4 Methodology

Regarding the used methodology in this project, it is necessary to differentiate, on the one hand, methodology and tools, and on the other hand, which way of work will be validated during its process

4.1 Chosen methodology

To perform the project, a variant of agile methodologies will be applied. The main objective of this methodology is to work on constantly improving and iterating the system. Therefore, it does not attempt to implement a system in parts, to finally obtain a product over time. The method consists, from a set of requirements, in implementing them in different iterations and at the end of each iteration obtaining a presentable product. Therefore, the following iterations are executed in order to improve the existing system.

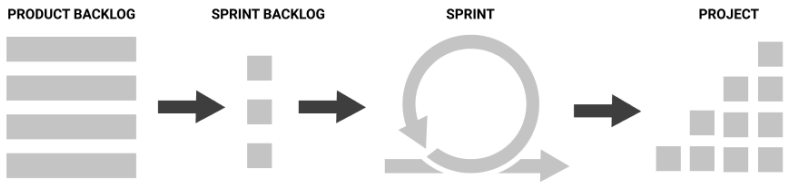


Figure 1: Structure of Agile Methodology

The reason why this method is going to be used is because given the three main requisites that there are: obtaining a set of indexed articles, implementing different techniques to classify the dataset and creating a web application; the project is going to focus on the second process. In this way, iterations will be adjusted to improve the classification techniques and to use as many methods as time permits.

In summary, from a system based on the three requirements implemented on the first iteration, a variation of the agile methodologies will be performed. Then, for each iteration, requirements will be done in order to improve and to expand the classification process.

This chosen method will provide significant flexibility when looking for ways to classify texts. The reason is that we do not attempt to classify with innovative and

efficient techniques, but experiment with different methods and depending on the level of learning difficulty and the results obtained will choose the next steps.

4.2 Risks

Theuse of all methods includes more or fewer risks that have to be detected in order to deal with them. About the agile methodology contingencies are related to productivity, knowledge and time. And in this section how to handle these risks will be managed.

The part that has to have more importance is the productivity level. One of the keys of iterate in agile methodologies is to, at the end of each iteration, evaluate the balance between effort and results. In the case that this fact is unbalanced, then it allows adapting the effort on the next iterations. For that reason, the project needs enough iterations to have this balance as much controlled as possible.

The other part, to have in mind, is related to the knowledge because depending on the developer that is working on a task will implement more or less with the same effort. The positive part is that the objectives of the project are adapted from the knowledge of one person only, so this problem would not appear.

Finally, another part that in this project is very established and can't be adapted is the time. The time is the first factor to define the project requisites but, in this case, is not a very problematic situation. The reason is that the project is oriented to do a small prototype to, then, iterate over improve the different classification models but, the final structure of the system will exist from the start.

4.3 Alterations

Regarding to the use of the chosen methodologies, the different modifications have been made:

Different iteration process

The original purpose was to implement on the first iteration a system prototype that included generating the dataset,a simple classification and the web service, for then improve the classification process. But finally, given that obtaining the dataset lasted longer than expected, and it took longer to implement a first simple classification due to lack of knowledge on the subject, it was decided to delay the

implementation of the web service and decide to do it as the last process in case of having enough time.

Therefore, the final methodology used was based on a project carried out in three phases, where the first and third consisted of satisfying the defined requirements, and the second in applying the agile methodology. It cannot be considered that the agile methodology has been used in all the processes of the project because, from the beginning, the project has not existed until its final date.

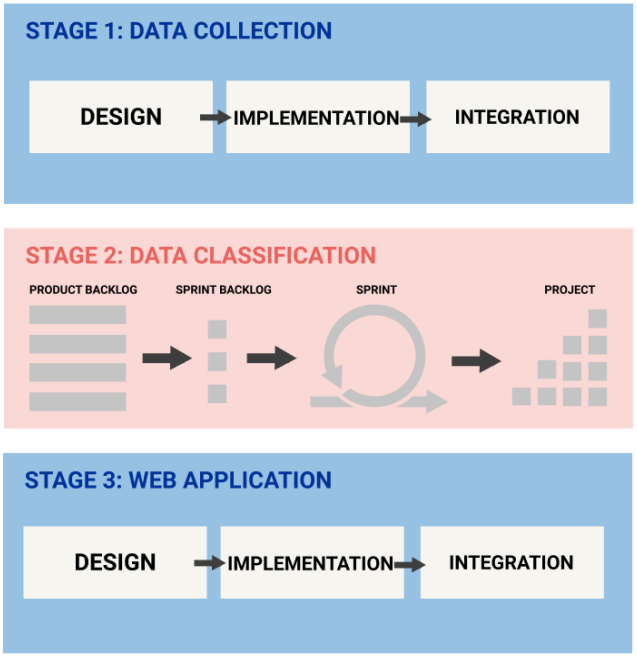


Figure 2: Final process of the project

5 Design and implementation

This section shows the overall architecture of the project applying the defined requisites, furthermore, the explanation of the techniques and methods used in it.

5.1 Architecture

For the architecture design to perform the experiments, all the text processes and classifications that require a specific implementation, further than use third-party libraries as sklearn or gensim, will be introduced into the system component called 'core'. So then, in the future web application, this part can be reused.

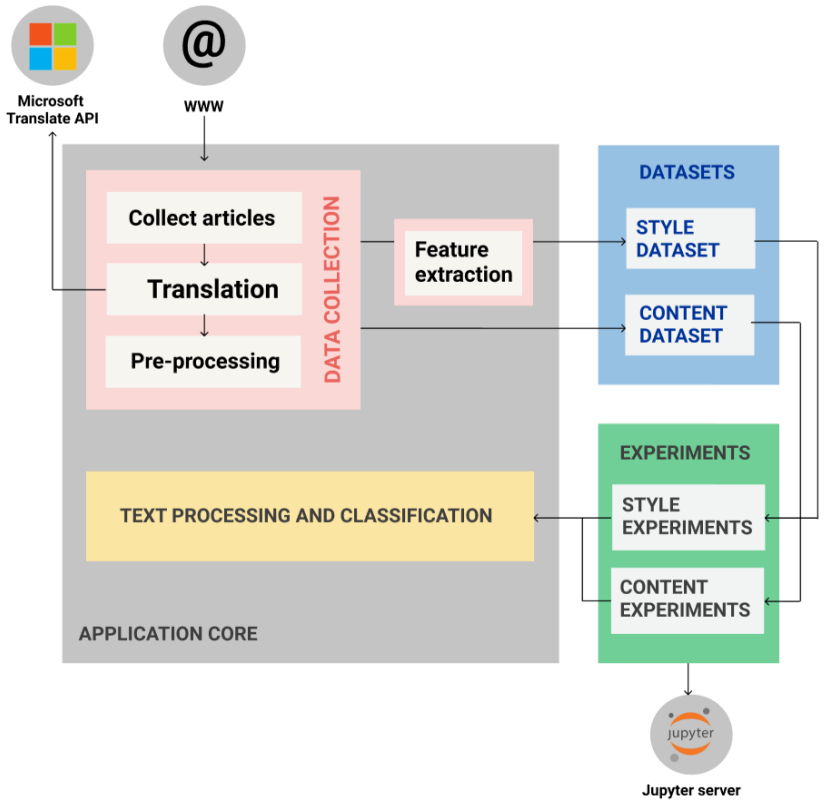


Figure 3: Architechture of fake news classificator

Apart from the core component, the central hub of all processes, the figure represents the structure of the set of pre-processed articles, of two different ways,

and also two different experiment focus. These experiments will be implemented in Jupyter Notebook, and for that reason, it will need a local server to run them.

5.2 Theoretical methods

The different techniques applied in the process of analysis and classification of documents are developed below.

5.2.1 Binary classification with SVM

Starting from the idea of Support Vector which refers to a vector of coordinates of an individual observation in space, we talk about Support Vector Machine as the algorithm that generates a boundary over the individuals we want to group [15].

Support Vector Machine is a supervised learning algorithm aimed at solving problems of classification and linear regression. In this case, the Support Vector Machine technique oriented to binary classification problems will be explained in more detail.

For binary classification, it is said that given a set of individuals placed in an n-dimensional space, the objective of the algorithm is to find the hyper-plane that maximises the separation between two classes.

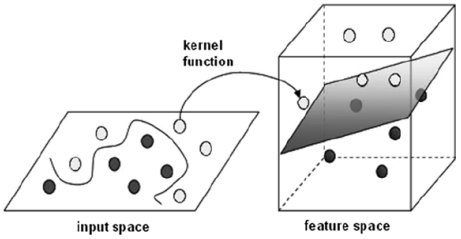


Figure 4: Representation of SVM distribution

There are different grouping techniques, also called kernels, which given their defined form in a function are intended to maximise the distance between classes. These kernels include the linear, polynomial, RBF and sigmoid models.

The positive aspect of this technique is that usually has good accuracy on the training process, and works good with a small set of data, as will be in this project. But also, in the case that individuals do not follow a grouping pattern the algorithm does not provide good results in the prediction.

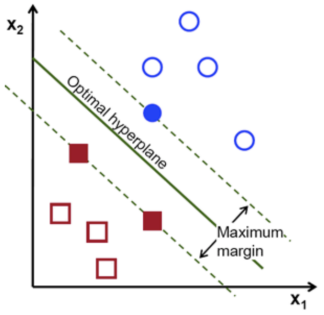


Figure 5: Representation of SVM borders parameters

5.2.2 Dimensional reductions

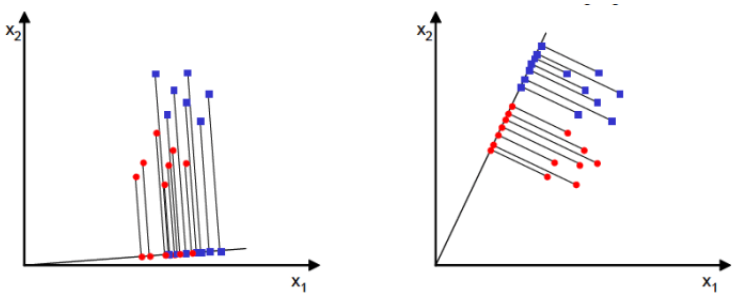
When talking about techniques of reduction of dimensions they are algorithms that have the objective of representing a set of data of a certain dimension to smaller dimension. These techniques try to simplify the data variables by means of grouping techniques but at the same time. They try to maintain the maximum possible information that the data have in the initial dimension. These methods different applications, one is the binary classification that will be dealt with more thoroughly. Two of the most popular methods are Principal Component Analysis and Linear Discriminant Analysis and they will be explained below.

- Principal Component Analysis

  The Principal Component Analysis(PCA) is a orthogonal transformation method that have the objective of convert a set of observations in linear correlation values, defined as *principal components*.[10] This is a unsupervised method because does not need to which class belongs each individual, the method tries to group the data by maximising the variance in them correlations.

- Linear Discriminant Analysis

  Linear Discriminant Analysis also starts from the idea that information can be represented linearly.[18] In order to do this, a reduction is made from dimension to dimension in which the information is projected onto a hyperplane of a lower dimension, until it reaches the pre-established dimension. It is also a supervised method since in order to carry out each reduction it is necessary to know which class each individual belongs to in order to maximise the distance between the two groups.



There are infinite possibilities to make the projection in the next smaller dimension, that's why, and as shown in the previous figure we try to maximize the separation between classes to be able to make a correct projection.

5.2.3 Natural Language Processing

Natural language processing is part of a branch of artificial intelligence that aims to process and understand human language[9]. This field can perform many different techniques depending on the application being used.

In this case, the information explained below is focused on text pre-processing techniques, vector transformations and topic modelling.

5.2.4 Preprocessing

Preprocessing techniques are all those techniques that are applied to eliminate text properties that may provide noise about future processing, and get a prepared structure to apply the final objective.

- **Noise removal**

  Depending on the origin of the document to be processed, it may be necessary to perform a series of steps to remove all those symbols and words that are not part of the implicit content of the text.

  An example can be found when it comes to processing the content of web pages that many times HTML tags and undesired symbols have to be removed as they are not part of the text.

- **Tokenisation**

  The process of tokenisation consists of decomposing a set of text into a sequence of elements called tokens, which are equivalent to the minimum

```
Input:    "<h1>Title</h1>
           <h2>Subtitle</h2>"
Output:   Title Subtitle
```

Table 1: Example of noise removal in text from website content

unit. These tokens can be both words and sets of words or symbols.

There are different strategies when it is tokenising, one of them is tokenisation by TreeBank. This strategy divides the whole text by words and symbols, with the exception that in the case of verbs contractions are kept in the same token. An example can be seen below:

```
Input:    "They'll save and invest more."
Output:   ['They', "'ll", 'save', 'and', 'invest', 'more', '.']
```

Table 2: Example of tokenization process with Treebank strategy

- **Standardisation**

  This process is usually done after the tokenisation process and aims to get the most out of the had words. Some of the most used techniques are the following:

  Removal of punctuation marks

  Most of the time, depending on how the text will be used, the set of punctuation marks may not provide useful information to the dataset. That is why there are cases where you perform this deletion.

  Modify upper letters

  Two equal words with the same semantic meaning can be written differently because one of them is at the beginning of the sentence.

  One of the proposed solutions to the problem is to transform all the letters into the same case. Everything and so there is the possibility that the two words have to have a different case because they have a different meaning. It will depend on the problem to be dealt with, which transformation is carried out in one way or another.

  Elimination of words known as 'stopwords'

  The `stopwords` are those words that regardless of the context are usually found in most text, and therefore does not provide any information to the

Input:	"All dogs are brown and all cats are black"
Output:	['All', 'dogs', 'are', 'brown', 'and', 'all', 'cats', 'black']
Output with lower case:	['all', 'dogs', 'are', 'brown', 'and', 'cats', 'black']

Table 3: Example of change words to lower case

properties are wanted to extract, so in certain cases it decides to discard them.

- **Lemmatization and steming**

  For grammatical reasons, words with a very similar meaning are often syntactic derivations. These derivations may simply consist in same words that only differ in gender or quantity. In some problems, it is convenient to group the words by their roots to reduce the grammatical diversity on documents. This is the main objective of the lemmatisation and stemming techniques [14].

  The stemming process consists of removing prefixes and suffixes from words to keep only the root. This process can offer some limitations since there are cases in which the root obtained is not an existing word.

Word	Deletion	Stemmed word
studying	-ing	study
studies	-es	studi

Table 4: Examples stemming process

  In the lemmatisation process, takes into account, the morphological analysis of words and requires a dictionary to save each equivalent root to perform the process correctly. Therefore, it may also be the case that the dictionary being consulted does not contain the transformation of a specific word. Everything and so it is usually the most used technique as it obtains the most correct results.

5.2.5 Vector transformation with TF-IDF

TF-IDF is part of the acronym Term Frequency-Inverse Document Frequency, is a word indexing strategy to evaluate the relevance of a word based on the calcu-

lated statistical weight. This algorithm calculates the weights of each word in the following way:

$$Tf - idf(w) = tf(w) * idf(w) \tag{1}$$

where:

tf(w) (Term Frequency) = Number of w repetitions in corpus / Total words in corpus

idf(w) (Inverse Document Frequency) = log(total number of documents / Documents that contains w)

This technique has multiple applications that work according to the relevance obtained from each word. Some examples consist of summarising documents according to the most relevant words or estimating the words considered 'stopwords' of a given domain, among many others.

5.2.6 Topic modelling with Latent Dirichlet Allocation

The Latent Dirichlet Allocation is described as a generative probabilistic model for collections of discrete data such text corpora [7] The method have the strategy of group a collection of documents by topics. This topics are defined by the probability of words that can be in the same document. Also this tool can give the topic distribution of a document to know the portion of each topic that a documents has.

The applications where this method can be used is in document modelling, text classification and collaborative filtering. About the text classification the process to implement consist on from a defined set of topics gets the topic distribution of each document in order to classify by the behaviour of each document.

5.3 Tools

In order to implement the project, the following third-party tools have been used:

5.3.1 Scikit-learn and Genism

There exist multiple libraries that have implemented different classification models or NLP algorithms, two examples are 'sklearn' and 'Gensim'. The decision to use

this tools it is to avoid implementing all the algorithms from scratch, and that these libraries have implemented. These open-source libraries are developed with Python, and with its use, more methods of classification could be applied to the project, taking care of the disponible time.

The library Scikit-learn includes a set of classification, regression and analysis algorithms. Moreover, it allows operating with datasets and library objects such as NumPy and SpiCy.

Por the other hand, the library Gesim, it is also an open-source library, but in this case, it focuses more text processing as topic modelling or words embeddings.

5.3.2 Jupyter notebook

The web application Jupyter Notebook is also an open-source platform, which it allows you to create documents with live code written in Python and run it inside. So then, this program allows you to write about the executed operations.

Jupyter Notebook is often used for multiple application such as creating statistical models, performing text modelling, using machine learning methods and so on. That is the reason why experiments will be executed in this program, and it also allows you to export the results easily.

6 Dataset

The dataset that is gonna be used is defined as the set of newspaper articles previously classified as true or false. To find a dataset that adapts to the project necessities, two proposals will be described as follows: Searching for an existent third-party dataset or creating a dataset of its own. In this way, the positive and negative parts of each option and finally the decision taken, will be evaluated.

6.1 Dataset selection

To be able to choose the dataset that will be used in this project, the properties that the dataset should have will be defined first. For that reason, the search of articles will be based on the following requirements:

1. Accurate data: Make sure that the items are correctly labeled as true or false.
2. License-free: Be able to use the dataset within the law.
3. Diversity of information: The dataset has to reflect the different topics that are currently covered in journalism. In this way, to promote the future re-training of the system without any limitation.
4. Reliable newspaper articles: Make sure that the original websites belong to serious companies.
5. Enough articles: Have a sufficient number of articles in order to train a system and to validate it.
6. Minimum knowledge about the content of the articles.

6.1.1 Option 1: Search for existing datasets

There are some open source platforms which their datasets can be used under a free license. Some examples are kaggle.com and github.com. After the research, the following two datasets are highlighted as interesting, and its negative and positive aspects will be studied.

Dataset 1: BuzzFeed. Top 100 fake viral articles from 2016 and 2017

Source

Firstly a collection of articles from an important social media company can be found. This company is **BuzzFeed** and its collection contains a set of the 50th fake articles more viral of the 2016 and 2017 which this sums 100 documents in total.

Each article contains the title of it and its URL. Then, in the case of use this dataset, the following tasks have to be done for obtaining a complete dataset:

1. To scrap each article for collecting all the text from the given URL.
2. From the same web pages that fake news were collected, also collect real ones with a similar style of documents.

So finally, about to use this dataset the following conclusions can be extracted:

Pros and cons of an existing dataset

Dataset	BuzzFeed
Pros	• Variety of topics • Free use
Cons	• The content of the articles have to be collected by URL • It only contains fake news. A search must be done to pick up a sample of the same size from real items. • Contains articles from not well-known websites • BuzzFeed is not officially validated by IFNC, still having a certain guarantee.

Dataset 2: Politifact. 240 articles about US politics

Source

This dataset, made by the organisation for fact-checking **PolitiFact**, is the second option to evaluate. This web site focuses its effort on check North American political news. The organisation is officially validated by the IFCN, so we have a complete guarantee from the reliability of the articles.

The mentioned dataset contains 120 real and 120 fake articles. For each article, not only the `title` and `text` can be found, it also includes other types of information such as the authors or the publication date. In summary, this option consists of a comprehensive and complete collection of North American political news.

With all this information, the following conclusions can be taken:

Pros and cons from an existing dataset

Dataset	PolitiFact
Pros	• Officially validated articles • The content of the articles is still available • Free use
Cons	• There are quite a few articles from unreliable sources • It contains articles of a very particular topic • Non-knowledge of political content

6.1.2 Option 2: Generate a dataset

The second direction for getting a set of classified articles, in order to work on the project, is looking up the Spanish website, **Maldita.es**. This organisation, validated by the IFCN, works on disproving fake news that are viralized in the Spanish network. In this way, the next option for getting a dataset is gonna be evaluated:

Dataset 3: Maldita.es. Set of disproved Spanish articles

In this website different types of hoaxes can be found. One kind is all the viral media in image or text messages format that any person could receive in its mobile phone. The other type, the one that the project is working on, is about all the articles published on the media. Even though the proportion of the second group is lower, an adequate number of articles can be collected.

To develop the dataset by consulting Maldita.es, the URLs of the fake news have to be collected manually. After that, a script for getting the content of those websites will be executed. And same as in the case of the Dataset 1 of BuzzFeed, the process for getting real articles will be done by getting the same number of articles from the same website of fake news.

On the basis of the issues raised, this option is assessed as follows:

Pros and cons from a created dataset

Dataset	Maldita.es
Pros	• Officially validated articles • Variety of topics • Free use • They come from medium-serious newspapers • Knowledge about the content of articles
Cons	• The content of the articles have to be collected by URL • It only contains fake news. A search must be done to pick up a sample of the same size from real items • They are articles in Spanish

6.1.3 Arised problems

After evaluating the different options encountered, and before making a decision, the following drawbacks encountered during this process should be highlighted:

1. **Trouble in finding complete and well-classified datasets.**

   The fact that fake news is a recent problem could be the reason why finding correct classified and free articles have been difficult.

2. **Most of the false information, which becomes viral in networks, are not newspaper articles.**

   One interesting fact found in `PolitiFact` and `Maldita` fact-checking web pages is that even they have different content from different countries, most of the disproved information is not from articles. Most of the viral news that disinforms are montages of images, audio messages or text messages that are expanded through social networks inconspicuously.

6.1.4 Conclusions

After evaluating the different mentioned options, all three cases have in common the fact of having the free use policy. The use of the data will be valid as long as its content is not distributed in a public source. Another common point is the similarity between the total amount of articles.

After the above reasons, **PolitiFact** dataset is the first discarded option. The fact that its content is limited to North American policy concerns is the main reason. Having more diverse articles could provide better results in our case study and the level at which the project aspires. In addition, it also could limit the use of the future web service to very specific articles.

When comparing the **BuzzFeed** and **Maldita** datasets, it can be seen that very similar steps have been taken. These steps consist of collecting the web content from its URL through a script, and from the content, filter the needed information, which in this case would be title, subtitle and text. Also adding the task of creating a list of URLs of certain news coming from similar websites.

Finally, the chosen dataset is from **Maldita**. The fact that, in contrast to BuzzFeed, the organisation is supported by the IFNC structure, and it provides veracity in the classification of articles. Even having to extract the URLs of the fake news manually, during the implementation of the project could be possible to continue expanding the dataset with new queries in this website.

6.2 Selected dataset

In this section, the process to do in order to collect fake and real news will be explained. Always with the premise that each fake news has to be refused on `Maldita.es` website.

6.2.1 List of articles

After looking for the entries on the website, where journalistic articles refused are contrasted, 70 web addresses from 20 different websites are obtained. Some examples of sites are elmundo.com, elpais.com or lavanguardia.com.

Given the list of websites about fake articles, the next step is to search for real articles. This search will be done manually and, to ensure the veracity of the article, it will be checked that the information is distributed in multiple media and, in the next days, has not been refused.

6.2.2 Process of content collection

The next step, in order to finally get the content of the articles list, is to generate a script that given the list of URLs returns the title, subtitle and text of each article.

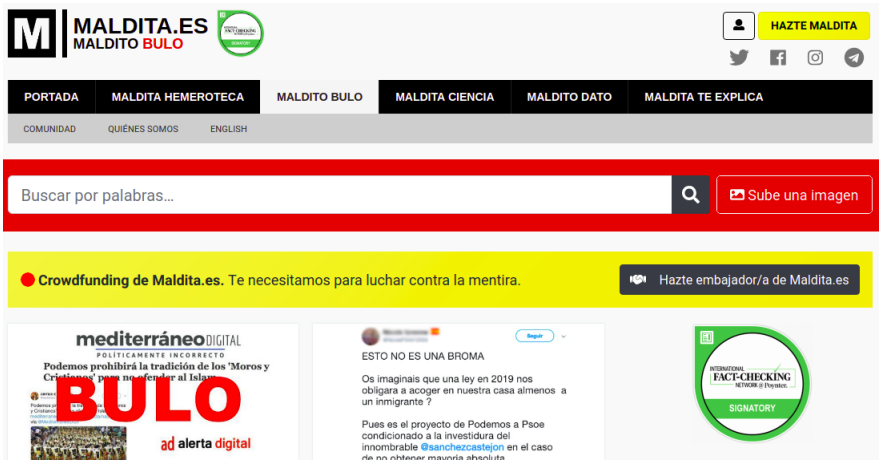


Figure 6: Maldita.es web site of fact-checking Spanish news

Once the necessary information has been obtained from each article, in order to be able to use the different word processing libraries, the content must be translated to English. Although the translation is not completely reliable, as in the experimentation no syntax analysis processes will be performed, this option was finally chosen. Therefore, the results will be affected but not enough.

The tool chosen for the translation is Microsoft Translator, which offers a free plan and provides acceptable results. After implementing a script to make calls to the Microsoft Text Translation API in order to translate the text, it is executed, and texts are stored in new files. So finally, each file makes reference to an article where includes the title, subtitle and text in English, and also the URL and its label.

6.2.3 Exploration

After carrying out the explained steps, a small exploration is performed in order to know better the data with it is going to work. It was observed that the dataset is composed by 137 articles in total where 70 are fake and 67 true news.

7 Analysis and classification based on the style of the articles

7.1 Objectives of the experiment

The main objectives of these experiments are the following:

- Extract quantifiable style properties from the dataset
- Explore properties and search for the most correlated variables
- Modelling a binary classifier using the SVM algorithm.
- Perform a dimensionality reduction with PCA and LDA to then classify the data entry with the SVM algorithm.
- Evaluate which reduction of dimensionality gives better results and if the classifier works better with this previous process.

7.2 Implementation

This section will explain each process carried out for the analysis and classification based on the style of the text. These processes have been defined based on the objectives of the experiment.

7.2.1 Data extraction

For classifying a text according to the style, it is necessary to extract quantitative properties that indicate how a text is written.

In this way the data, which has been handled with the mentioned techniques, will not be fragments of text, it will be numerical variables extracted from each document. These variables are part of the following style aspects from a text: quantified data, complexity and sentiment.

Label:	
`fake`	True if the document is fake, False otherwise
Quantity:	

n_words	Total number of words
n_sentences	Total number of sentences
pert_total_adj	Percentage of total adjectives
pert_total_conj_prep	Percentage of total conjunctions and prepositions
pert_total_verbs	Percentage of total verbs
pert_total_nouns	Percentage of total nouns
title_n_words	Total number of title words
title_pert_total_conj_prep	Total percentage of conjunctions and prepositions in title
Complexity:	
mean_character_per_word	Mean number character for each word
mean_noun_phrases	Mean number of noun phrases in document
mean_words_per_sentence	Mean number of words for each sentence
pert_different_words	Percentage of different words
Sentiment:	
pert_total_negative_words	Percentage of negative words
pert_total_positive_words	Percentage of positive words
sentiment	Sentiment score
title_pert_total_negative_words	Percentage of negative words of the title
title_pert_total_positive_words	Percentage of positive words of the title
title_sentiment	Sentiment score of the title

Table 8: Variables of style dataset

Before calculating these characteristics, a natural language preprocessing will be applied to minimise the noise that could arise from words or symbols which are not technically part of the text or do not help in order to improve the results.

In some of the properties of this experiment, the documents need to be tokenized in one way or another. For this reason, the procedure to be done consists of, firstly, tokenizing the documents by sentences and then preprocessing each sentence individually.

This structure allows to adapt the information according to the data that wants to be computed because an union of all sentences in words can be done without repeating the preprocessing.

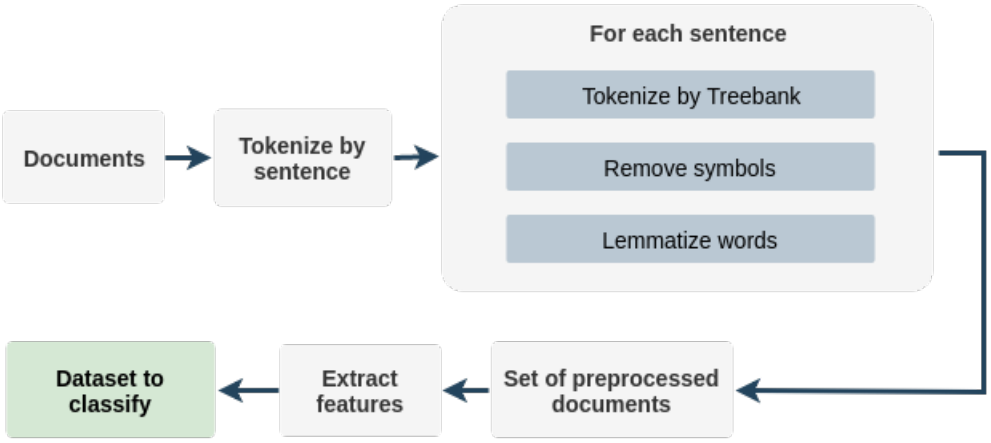


Figure 7: Text preprocessing for style classification

Regarding the preprocessing of each sentence, the content is tokenised by the Treebank strategy. Then all the contractions and punctuation marks are deleted because they do not provide useful information for the use case.

In the next step, the words are lemmatised in order to standardise them. Finally, the first letter of those words, which has been found at the beginning of the sentence, is transformed to lower case. It is known that is not the most optimal process, but it is considered as the best solution given the results and the spent time. The reason is that it is working with documents that contain many proper names and it wants to avoid transforming everything to lower case.

In this case, it is not interesting to remove stop words because they will help to quantify some of the mentioned properties, such as the percentage of conjunctions. Finally, once this process is done, it will be saved as a new dataset to be imported.

7.2.2 Data exploration

The objective of the data exploration is to check the correlation between variables, especially between output variables. Studying the data will allow to know about which variables could get better results on the classification process.

Taking care that the correlation value is inside the [-1, 1] interval, the numeric relation that exists between two variables. So, when the value is closer to zero, a lower correlation can be observed with the calculated variables.

After computing the correlation values from the output variable, which is called

fake, with the rest, more than expected values closer to zero are obtained. Although, the most correlated variable with *fake* is `pert_different_words` with *0.309* as score. Then, after sorting the rest of the correlations by value, the following list is obtained with scores greater than 0.2:

Top correlations with 'fake'	Correlation score
`pert_different_words`	0.309559
`n_words`	0.270078
`pert_total_nouns`	0.263984
`title_sentiment`	0.230889
`mean_character_per_word`	0.228940
`pert_total_negative_words`	0.227759
`pert_total_verbs`	0.225057
`pert_total_adj`	0.211848

Table 9: Variables more correlated with the label fake

This punctuation shows that some difficulty in processing fake news could exist. On the worked data, it is not possible to observe any direct relation. However, the process is repeated in order to observe the correlations between themselves from this subset with the objective of discarding the repeated information that two different variables could give.

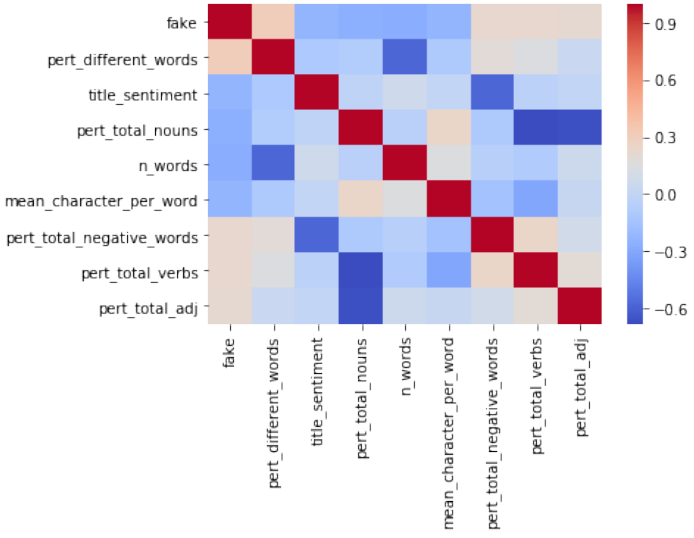


Figure 8: Map of more correlated features with the variable 'fake'

In this new correlation relationship, some high correlation values are distinguished.

One scenario is the set of three variables pert_total_nouns, pert_total_adj and pert_total_verbs. By interpreting the fact, their correlations have sense because they are portions and not absolute values, so when a variable is greater, the rest are minors as a consequence. Another observed correlated group is the formed by title_sentiment with pert_total_negative_words, where can also be assumed because both are dealing with the sentiment.

7.2.3 Training and validation datasets

Given the information obtained, in the first part of the experiments the most correlated values with fake label, and that also are not correlated between themselves, will be used. These variables will be the ones which will train and validate the different classifiers models.

Variables of dataset
fake
pert_different_words
n_words
pert_total_nouns
title_sentiment
mean_character_per_word

Table 10: More correlated variables with the label fake

During the training and validation process, splitting the sets into 80% for training and 20% for validation is decided. Also, the same seed will be used in order to perform the different methods with the same subsets of data.

Also, before the data is divided, each input variables will be standardised in order to obtain better results during the classification processes.

7.2.4 Direct classification with Support Vector Machine

On this section, the data will be classified training an SVM model without processing the data before. For creating the different SVM various a different bunch of kernels will be applied and each hyperplane parameter values will be optimised. Once the different models are created, they will be trained and validated to finally evaluate and compare the results.

For the model creation, the `GridSearchSVC` tool will be used in order to get the optimal values of `C` and `gamma` parameters. These parameters will adjust the boundary between both classes. The step is done for each kernel as it is showed below:

```
models['rbf'] = svm.SVC(kernel='rbf', C= 10, gamma=0.01)
models['linear']  = svm.SVC(kernel='linear', C= 0.1, gamma=0.001)
models['poly']  = svm.SVC(kernel='poly',C= 1, gamma=1)
models['sigmoid'] = svm.SVC(kernel='sigmoid', C= 10, gamma=0.01)
```

After creating the different SVM models, and with the training data, the training process of the model is executed. Then the rest of the data is validated for getting the model predictions. In order to evaluate these predictions the accuracy score is calculated. This accuracy refers to the good of bad classification of the model.

Kernel model	Train score	Validation score
rbf	0.77064	0.6
linear	0.74311	0.64285
poly	0.88073	0.72727
sigmoid	0.74311	0.64285

Table 11: SVM Accuracy scores by kernel

As seen at the first time with the low correlation between variables, the results show this situation. The training data has already showed that they are unable to find a pattern in order to classify if the documents are fake or not. Moreover, in the validation is reaffirmed. However, the polynomial SVM model has better accuracy in the training and the validating aspects. For that reason, the polynomial kernel is the most adapted to the data.

7.2.5 Reduced dimensions with Principal Component Analysis and Linear Discriminant Analysis

The next method for classifying the dataset by its style that will be used is applying different dimensional reduction techniques. These techniques have the objective of reducing the number of dimensions from a certain variable, keeping as much information as possible.

PCA dimensional reduction

```
from sklearn.decomposition import PCA
pca_model = PCA(n_components=2) # Create model
```

LDA dimensional reduction

```
from sklearn.discriminant_analysis
        import LinearDiscriminantAnalysis as LDA
lda_model = LDA(n_components = 20)
X_lda_train=lda_model.fit(X_std_train, Y_train).transform(X_std_train)
X_lda_test=lda_model.transform(X_std_test)
```

7.2.6 Classification with reduced dimension data

Once the data dimensionality are reduced by the different methods, the same classification process, from the last experiment, is applied with the objective of seeing the results but also for comparing both used strategies. Given that objective, from the output of the reduction process, the variables [x,y] are classified and these results are obtained:

Classification from PCA reduction

Kernel model	Train score	Validation score
rbf	0.69724	0.63636
linear	0.70642	0.58333
poly	0.68807	0.875
sigmoid	0.66055	0.5

Classification from LDA reduction

Kernel model	Train score	Validation score
rbf	0.76146	0.64285
linear	0.76146	0.64285
poly	0.63302	0.52631
sigmoid	0.76146	0.64285

In both used methods, the training results are lower than the classified data without the reduction process. That also returns a lower accuracy score on the data validation as a consequence. So the results, in this case, are not significant.

Moreover, the classification with a polynomial kernel after a PCA performing is an exception. In this case, the accuracy score in the validation process is 0.87. As it is shown below, the model in the training process has not been able of grouping

the data in two classes perfectly, even though the validation score is good enough, where part of the probability work has influenced in the results. Since the model was not able to group the initial data, a good classification cannot be considered.

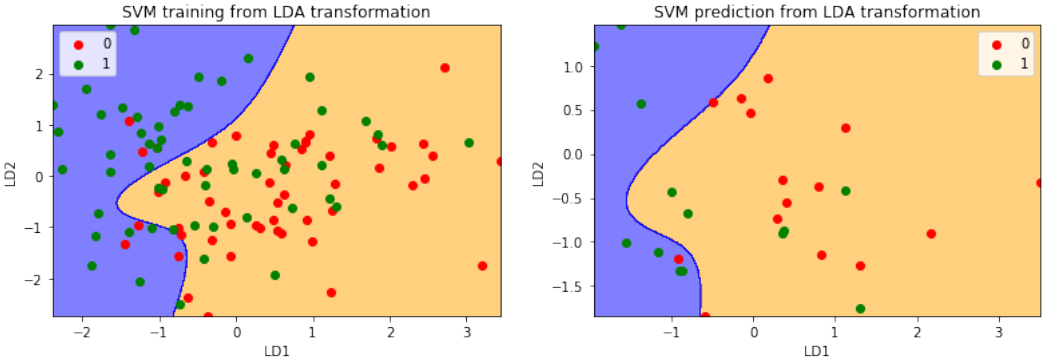


Figure 9: SVM poly classification from LDA transformation

8 Analysis and classification based on the content of the articles

8.1 Experiment objectives

The main objectives of these experiments are the following:

1. Transform text with preprocessing natural language techniques.
2. Explore the content of the dataset that will be used and check the word distribution.
3. Perform an SVM classification from most TF-IDF relevant words and their similarity.
4. Perform an SVM classification from topic distribution made with LDA method.
5. Perform an SVM classification from most TF-IDF relevant words and their similarity from Doc2Vec space distribution.
6. Compare and evaluate the different results of each method.

8.2 Implementation

8.2.1 Data extraction

For the next group of experiments, a new text preprocessing is needed, in order to perform them. In this case, for analysing the articles by the text context, the text preprocessing will be different in comparison to the style experiments. The big differences are there will be only one variable and it will be the content of the article.

The documents text preprocessing will consist of the same experiments carried out previously with a few differences. Firstly, by the Treebank strategy, the data will be tokenized and the removed symbols for only working with words. Then, each token will be lemmatised to finally remove the considered stop words. The next figure represents the steps to perform:

After an applied preprocessing and with the cleaned text, all the set of articles will be stored in order to be explored and used on the following experiments.

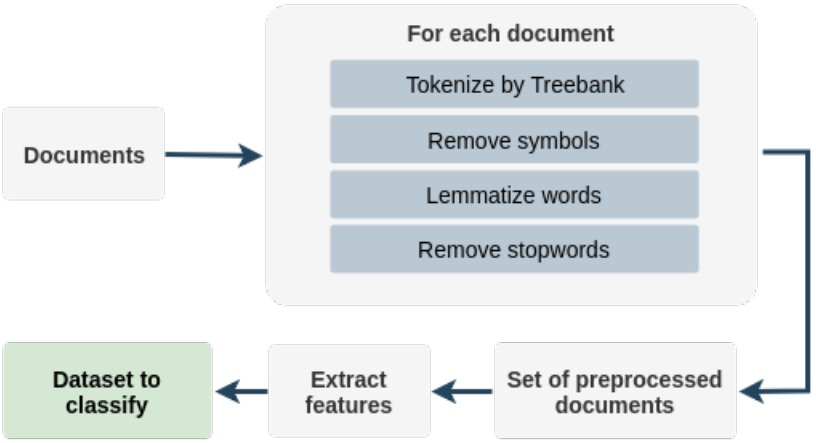


Figure 10: Text preprocessing for content classification

8.2.2 Data exploration

In order to know more about the content of the articles, an exploration will be done for observing the word differences between fake and real news and basically the similarity between both types of articles will be calculated.

Regarding the most repeated words in the corpus, the most used words depending on the type of each document and their distribution are represented in the next plots.

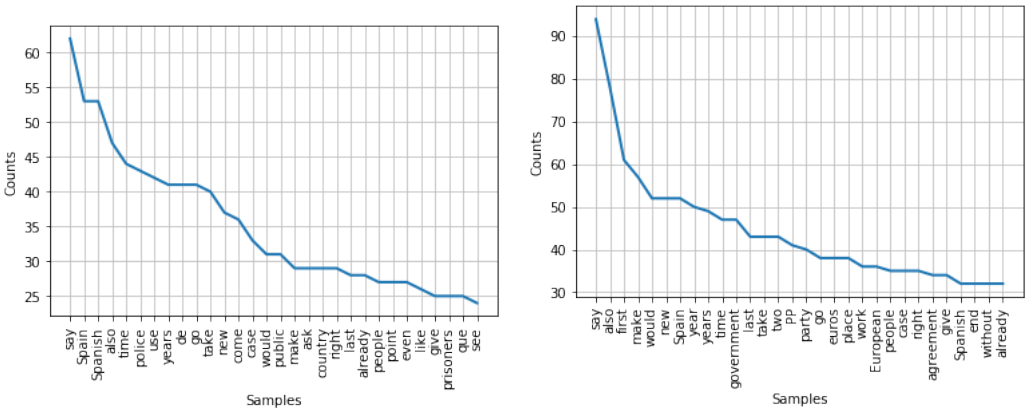


Figure 11: Word distribution by occurrences in dataset

At the generation level, it can be seen that the data being processed belongs to current news of the Spanish country. Another aspect to emphasise is that there are not many differences in the repeated words between both groups, therefore

in the following experiments, it will be avoided to use the repetition as a factor to classify.

8.2.3 Training and validation datasets

In order to classify the data and to do the experiments under similar conditions to those above, the dataset will be split in two with the same parameters as used on the style experiments. So, 80% of the dataset will be for the training process and the 20% for the validation. In order to get the same objects for different experiments, the same seed will be used for the partition.

8.2.4 Classification with TF-IDF and cosine similarity

For this classification, the TF-IDF method and the cosine-similarity calculation are going to be used. The main idea of this experiment is to create one document that includes the most relevant words for each type of articles. Then training an SVM model with the similarity, that each article has with both documents, as an input data. This idea is reproduced in the next figure:

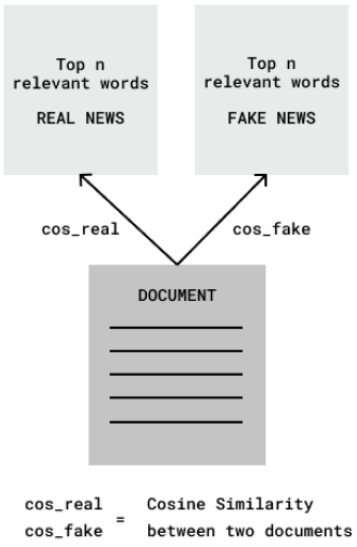


Figure 12: Classification strcuture by extracting the most relevant words with TF-IDF

First of all, a dictionary that represents each word as a number is created. For the experiment case, this dictionary will be created with the words from all the document dataset, so the same words will have the same representation. The next step is to get the documents that will be used in order to extract their most relevant words. These documents will be the real and the fake articles from the training dataset, so they will be grouped by its label.

```
cv = CountVectorizer()
X_train_counts = cv.fit_transform(dataset['text'].values)
# Split by type of documents
df_train_real = df_train.loc[df_train['label'] == 0]
df_train_fake = df_train.loc[df_train['label'] == 1]
cv_train_real = cv.transform(df_train_real['text'])
cv_train_fake = cv.transform(df_train_fake['text'])
```

Once the articles have been grouped by type, each method for getting the TF-IDF distribution is applied in order to get a relevance weight for each word. Then, each list of words will be sorted by their relevant weight and by the 600 most relevant words. In the next table, it is possible to observe the ten most relevant words for real and fake articles from the training dataset.

Word	Relevance score
rice	0.689
cheese	0.637
dog	0.624
restaurant	0.5
ikea	0.5
switzerland	0.481
crocodiles	0.48
sexual	0.458
foreing	0.442
songs	0.458

Table 14: Most relevant works from fake news

Word	Relevance score
bitcoin	0.766
mw	0.578
columbus	0.495
attack	0.478

maroto	0.442
education	0.437
degree	0.423
valeria	0.418
burst	0.414
passengers	0.409

Table 15: Most relevant works from real news

It is possible to observe that the ten most relevant words of each group haven't got a semantic relation between them and they are not included in both lists. Also, exploring each set of 600 words, it is possible to observe that 142 terms are repeated in both documents, which is the 25% of all the unique words.

After a brief exploration of the words representatives from fake and real news, the similarity, of each article with these two new documents, is calculated in order to train an SVM model with this data.

```
for index, row in df_train.iterrows():
    to_number = cv_model.transform([row['text']])
    cosine_sim_fake = t.get_cosine_similarity(cv_top_fake_words,
    to_number)
    cosine_sim_real = t.get_cosine_similarity(cv_top_real_words,
    to_number)
    dataset.at[index,'cos_fake'] = cosine_sim_fake[0]
    dataset.at[index,'cos_real'] = cosine_sim_real[0]
```

When the similarities have been calculated, and in the same way that the lasts experiments, the training and prediction processes are done for different SVM kernels with optimal parameters. So finally, after completing the classification of all documents about the similarity of the 600 most relevant words of each type of documents, the following results are obtained:

Kernel model	Train score	Validation score
rbf	0.98165	0.75
linear	0.98165	0.75
poly	0.52293	0.46428
sigmoid	0.98165	0.75

Taking the results, it is possible to observe good results except those of the polynomial kernel. The polynomial kernel, with a 0.52293 of training score, was not

able to train the model in order to group each class together, in consequence, the validation score is also lower because any pattern was detected.

Otherwise, after an optimal training result, the remaining kernels have achieved acceptable validation results with a score of 0.75.

As this classification consists of working with two dimensions, it is possible to graphically observe each classification process of one of the models with better results. So the training and the validation process of the data is represented inside the RFB model grouping.

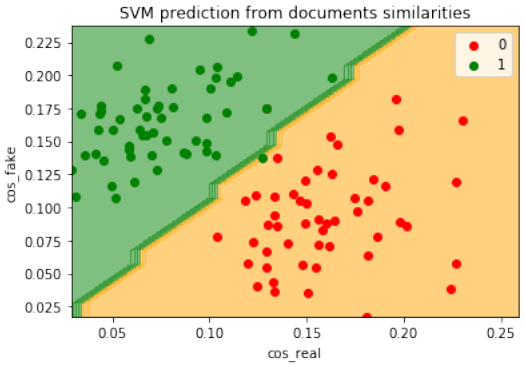


Figure 13: Train process of SVM rfb model

On the training process, when the model tries to trace the border for dividing the two classes of articles, an almost perfect division can be seen. Only one individual item from all the elements was not good classified.

In addition, to be able to see that the individuals of each type of article are keeping related values of similarity on the document of their type, there is also a lower significant similarity with the document of the class to the one which they do not belong. Therefore, for the first time so far, the clearest pattern has been found that distinguishes true news from false news.

Regarding the validation process, the results are more dispersed between the two similarities even though the 75% of the individuals are correctly classified.

One of the reasons why the result of the validation process is not optimal is because the calculated similarity has been made with the most relevant words of the training dataset. However, the objective was to observe whether the relevant words of the training dataset were also similar to those of the validation dataset. From the obtained results, it can be stated that this relationship exists with the most relevant words.

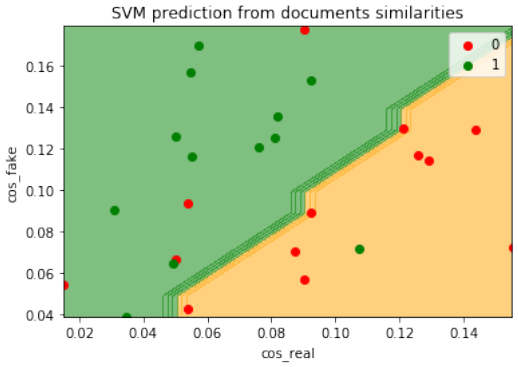


Figure 14: Validation process of SVM rfb model

The optimum value of the most relevant number of words

The previous classification has been carried out given a specific number of relevant words. The aim of the following process is to observe the optimal number of relevant words where the SVM classifier obtains the best validation result. For this purpose, a script will be generated and it will classify the documents according to the similarity with the set of most relevant words within the range [50, 4000] with a step of 50 words.

Once the script has been executed for classifying the similarities, the maximum score of the four created models in the training and in the validation processes is collected for each N-value of relevant words. For each N-value, the maximum score achieved in each process is shown below:

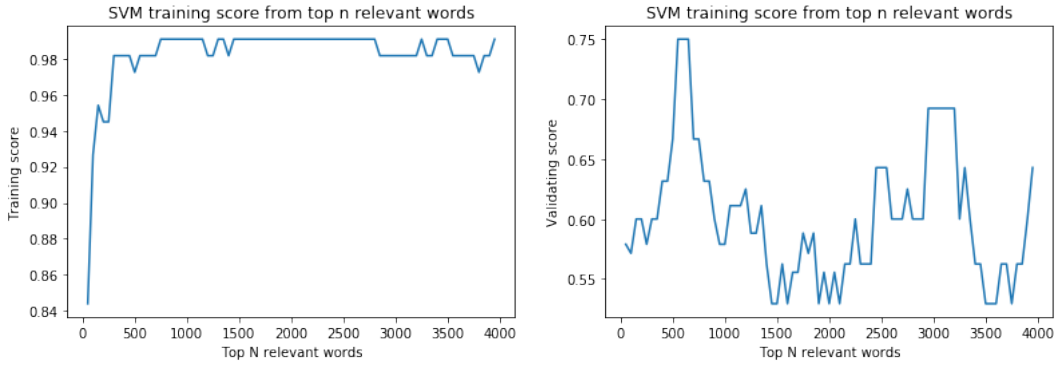


Figure 15: Main sctructure to classify from TF-IDF relevant words and cosine similarity

In the training process, it is observed that until a certain N-value is reached, the

results are not optimal. From this N-value the models are trained correctly with a score close to 0.98.

Respect to the scores of the validation process, interesting values to know if the data have followed a group-able pattern, very irregular results were observed as a function of N. These go from the accuracy of 0.75 to less than 0.55. On these results, it is possible to affirm that when N takes the value form 500 to 600 and 3000 two local maximums are observed. But in conclusion, the optimum classification value of this dataset is when N has a value within [500,600].

8.2.5 Classification from Latent Dirichlet Allocation topic distribution

From the method of Latent Dirichlet Allocation, it is possible to group the documents by topics and to know for each document what portion of each topic they have, which will be called topic distribution. Thanks to this technique, the objective will be bundle the training dataset in a different number of topics and will classify an SVM model from the topic distribution of each document.

For performing this classification, the optimal SVM model will be searched from a concrete number of topics and, then, the behaviour of SVM predictions, as a function of N, will be analysed.

In order to perform the first part of the experiment, the LDA model is created from the training dataset for detecting 20 topics. Some of the topics detected are showed below:

```
Topic: 0
Words: 0.005*"Vox" + 0.005*"degree" + 0.005*"Cs" + 0.003*"students" +
    0.003*"technical" + 0.003*"Barcelona" + 0.003*"PP" + 0.003*"Rivera"
     + 0.003*"title" + 0.003*"university"
Topic: 1
Words: 0.004*"yellow" + 0.004*"introduce" + 0.004*"products" + 0.004*"
    Chinese" + 0.004*"center" + 0.003*"brand" + 0.003*"fuel" + 0.003*"
    network" + 0.003*"Brussels" + 0.003*"illegally"
Topic: 2
Words: 0.006*"de" + 0.005*"young" + 0.005*"que" + 0.004*"politician" +
    0.003*"case" + 0.003*"Pedro" + 0.003*"woman" + 0.003*"would" +
    0.003*"Sanchez" + 0.003*"court"
```

Figure 16: Topic distribution of an example of LDA model

It can be observed, in the topics shown, the existence of a semantic relationship between the different words that makes up the topic. The clearest example is seen

in the ”Topic 0” where it includes words referring to different traditional Spanish parties and issues related to those parties.

On the next step, the topic distribution of each document from the training and the validation dataset is calculated and will be the input in order to train the different classifiers. From each kernel the following results of both processes are taken:

Kernel model	Train score	Validation score
rbf	0.80733	0.8
linear	0.77064	0.58823
poly	0.80733	1.0
sigmoid	0.77064	0.58823

Scattered results can be observed between the different kernels classifier in the training and the validation processes. Although, it got good enough results in some kernels, they are not reliable. If the process is repeated different times, very diverse results are obtained and, for example, the model with a RBF kernel can not get over 0.4 of accuracy. This fact happens when the topic modelling process. Each time executed provides very different themes so it is not possible to train an stable classification with this strategy.

The same happens when trying to find the optimal number of topics where the classifier provides the best results.

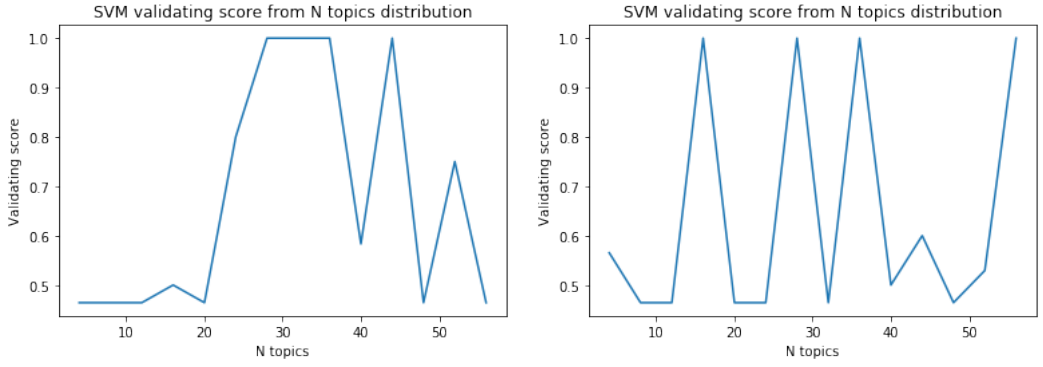


Figure 17: Main sctructure to classify from TF-IDF relevant words and cosine similarity

If the validation results of the optimum score are represented, for each N value of topics, a great irregularity is observed and it corresponds to the absence of a tendency. But when the experiment is repeated under the same conditions, the results are very different again.

Therefore, with this experiment, it can be concluded that given the set of data we have it, it is not possible to extract reliable subjects from the documents with the LDA tool and then classifying them according to the distribution of the documents on the subjects.

9 Web Service

After experiment with different classification strategies, the next step is to give usability to one of the used methods with a web service creation. In this section, the objectives, methods and the classifier included in this web service will be explained.

9.1 Introduction

The lack of the number of articles has suppose a not great results in the different performed experiments. From the rigid schedule that the project has, its scope was defined lower for the complete topic that it is working on. In view of this situation, the expected results made to focus the project by including a system that could improve the base created with the information that can be added during the time.

One of the tought solutions to improve this situation is creating a web service that not only was able to consult and predict and article from the internet but rather was able to improve the classifier re-training with new data. So on the next chapters, the design and implementation about this system is explained to understand how the web service can include this functionality.

9.2 Design

9.2.1 Objetives

After all the explained, the objectives of the web serice implementation can be summarized in the two next points:

1. Implement a system that can consult an article from the web and predict its reliability with the implemented classifier
2. Implement a system that in case of return a bad prediction from a consult, allow to inform the system to re-train the classifier.

9.2.2 Architecture

The structure of the program will consist in an API developed with Flask, a framework developed in Python, that it will have two serialised processed, the classifier creation and the API methods. The formed architecture it showed on the figure below:

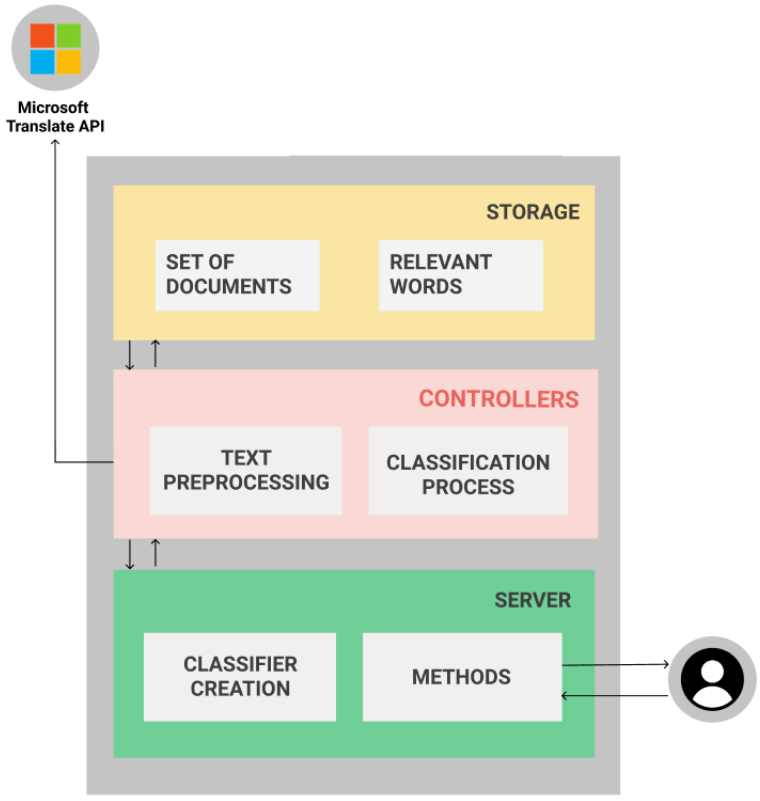


Figure 18: Web service architecture

On the first part of the system, when the program is initialised, the chosen classifier will be created and trained with all the processed included. This processes were done before for the experimentation cases by scripts, and now will be joined to be able to do in live. As can be seen, first the initial dataset will be read from the system in order to execute the same steps as done on the experimentation to final train the model. This model will be sent to the app to wait any method call of one of each consults.

The second part of the system consists in two different methods, one GET and one POST. The GET method will be the one to predicted a consulted article. The processes that this method have to include are similar to the dataset creation for the experiment, so most of the functions are rehoused from the dataset building.

The POST method consists on modify the classifier that predicted a bad classification of a consulted article, so this method will implement a procedure to include the last consulted article into the classifier with the defined label.

9.2.3 Classifier

The classifier that will be included in this system will be the implemented classifier explained on section 8.2.4. This classifier is the one that from the content of the articles classify by the similarity of the most relevant words of each type of articles.

The decision of choose this method is because was considered as important classify the articles from the content and not from the style. The style is something that can guide the people to suspect if the article is or not fake, but as seen on the introduction, fake new always try to move on the people and this characteristic is easier to find in the content and not in the style.

So, after decide use a content classifier the only which gives a good results was the mentioned. And for including the classifier into the system is only necessary to follow the same steps as the explained in the performed experiment.

9.3 Implementation

In this section will be included the folder structure of the system and describe the implemented methods in the web services.

9.3.1 Folder structure

The folder structure following on the implementation consist on the main file named `app.py` that runs the server and some folders grouped by its methods purposes.

The `app.py` will be the responsible on the classifier creation, when the Flask server is initialised, and also the runner of the two implemented methods. This

file will be connected with the controller, the file responsible of control all the prediction and training processes.

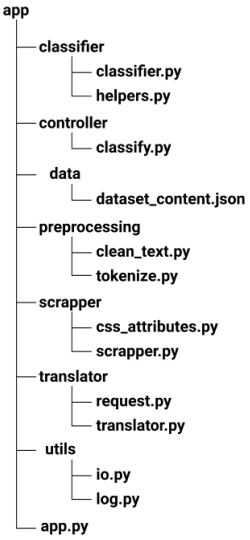


Figure 19: Web service architecture

The other folders included implement each functionality of the system. The `classifier` folder is the responsible to the classifier creation and its approach is to train and predict with the consulted articles. In the other hand, the `preprocessing` folder is the used on the classification process to clean all the documents that will be used by the classifier. And finally the `translator` and `scrapper` has similar approaches because they have implement different web request to translate or get the web content.

9.3.2 Methods

GET: `/predict/`

Predict an article from its URL

Headers	
url	URL of the article to consult
page	media company of consulted article

Responses:

POST: `/infer/`

Code	Reason	Message
200	Good response	Show the classifier prediction
400	Bad request.	The message indicate the reason of the error

Gets the last articles classified and re-train the classifier with the indicated labeled.

Headers	
label	Set true of false the consulted article.

Responses:

Code	Reason	Message
200	Good response	Good classifier training
400	Bad request.	The message indicate the reason of the error

9.4 Conclusions

From this implementation, the idea is observe the classifier prediction and see how it works and at the same time improve its prediction with new data included on the system.

The next step of the implementation was to calculate the level of improvement that the chosen classifier has done with new data. The limited time of the project didn't let do this process, but the pathway to do that in implemented for do it in future work.

10 Project planning

The project is estimated in an effort of 18 ECTS credits, of which 3 are part of the GEP course.

Each credit is estimated at 30 hours and the total estimated hours for the course would be equivalent to 90 hours, so 450 hours are assigned to the whole project.

In addition, the duration of the Final Project is estimated at 18 working weeks. That is why an effort of approximately 30 hours per week is calculated on average.

10.1 Schedule

10.1.1 Calendar

The table below shows the project deadlines, defined by the university, that has to be followed on the schedule:

	Terms
Start of the project	September 17, 2018
Start of GEP course	September 17, 2018
End of GEP course	October 22, 2018
Finall follow-up meeting	December 17, 2018
Oral defense of the project	January 31, 2019

Table 18: Calendar schedule of Final Degree Project

10.1.2 Tasks

In order to be able to plan the project process, the different objectives were divided into the following tasks so that most of them could be carried out sequentially and others in parallel.

GEP Course
`Expected time:  90 hours`

Realization of the course of management of projects with the objective of focusing, defining and planning the project to be able to realize it later.

Research
`Expected time: 45 hours`

Process of research on the state of the art and learning about the techniques to be used in order to achieve the objectives of the project.

Set-up
`Expected time: 10 hours`

Decide which tools will be used and configure the entire development environment prior to deployment

Defition of requeriments
`Expected time: 6 hours`

To define the initial requirements that the project must have in order to achieve its objectives in order to carry out the methodology used.

Implementation of the requeriments
`Expected time: 225 hours`

Perform project analysis and implementation based on project requirements from different previously established iterations. These iterations consist of a week and a half and a total of five will be carried out.

Analysis of results
`Expected time: 30 hours`

Once the implementation process is finished, the results obtained within the study will be analyzed.

Project conclusions
`Expected time: 30 hours`

With all obtained results, take into account if the project has approach them objectives.

Final documentation
`Expected time: 75 hours`

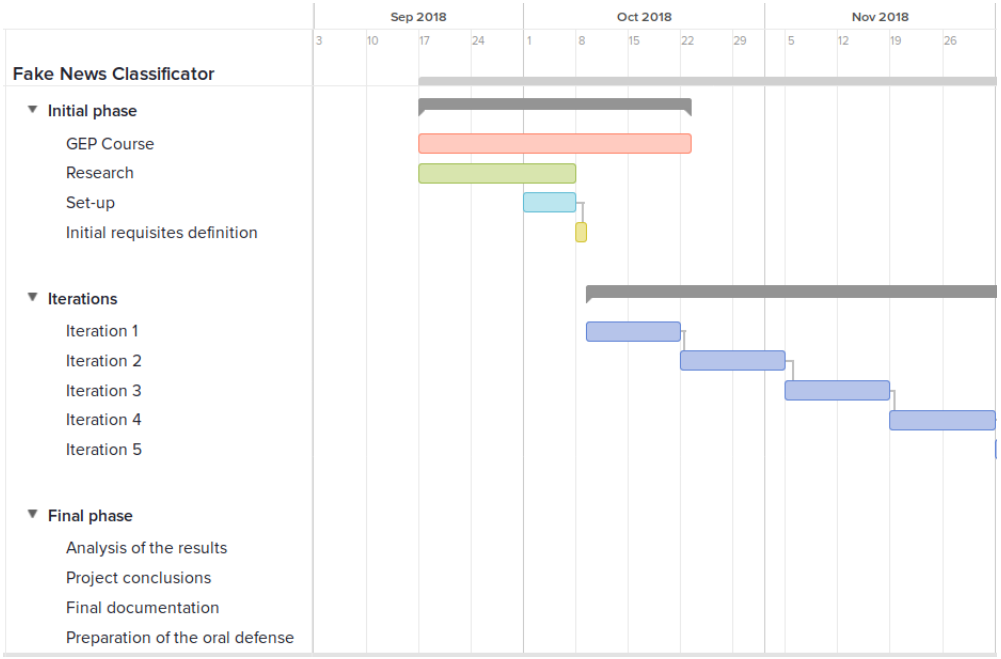
Once the system has been implemented and its functioning has been analysed, the entire project process and the final results will be documented in order to be able to deliver it.

Oral defense
`Expected time: 30 hours`

Finally, when defending in a non-native language, more time than usual will be devoted to the preparation of the oral defence.

10.1.3 GANTT Diagram

Based on the tasks, and taking into account the deadlines for each phase of the project, the following planning was established:



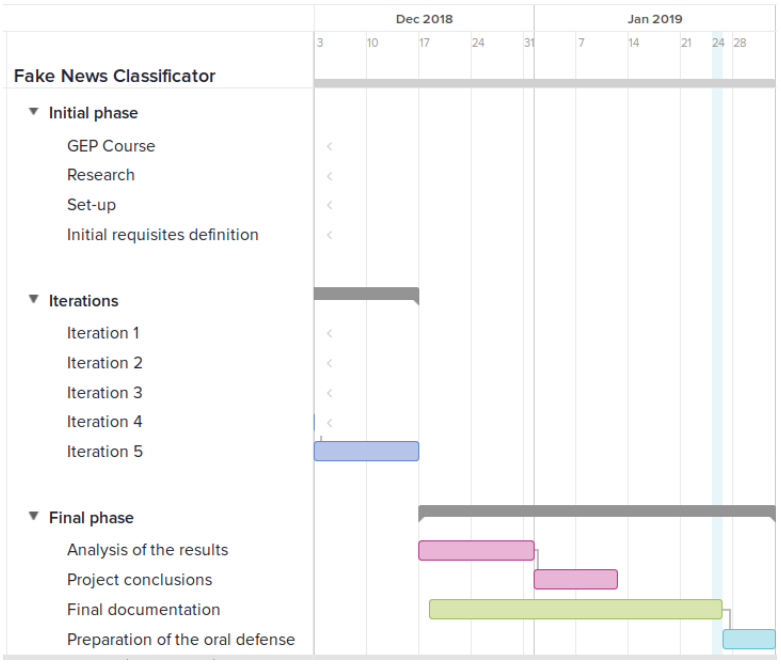


Figure 20: Original project schedule

10.2 Alternatives and action plan

During the project development process, it may appear problems that affect the previous planning. These alterations can derive different origin will be taken into account to complete the objectives of the project.

10.2.1 Learning process

During this project, the author will learn and implement a lot of new knowledge that has not been given in the degree. This means that the learning curve has to be taken into account and it is possible that it alters the effectiveness, especially of the implementation. In such a case, the methodology used already takes this aspect into account and in the case of having a long or short learning time, it will only influence to carry out more or fewer experiments, and the objectives of the project will continue to be achieved.

10.2.2 Instability in the effort of hours

Another possible problem arises when the project actors do not follow the defined planning, causing a delay in the execution of the project.In this case also, with the methodology put into practice, it provides that if the actors can not devote the same effort in each iteration and they commit to recover them in the following, no problem would end up arising. The reason is that the fulfilment of the requirements is not ruled by specific deadlines, since the only deadline is the final date of the project.

10.3 Changes from the initial planning

10.3.1 Delay on scheduling

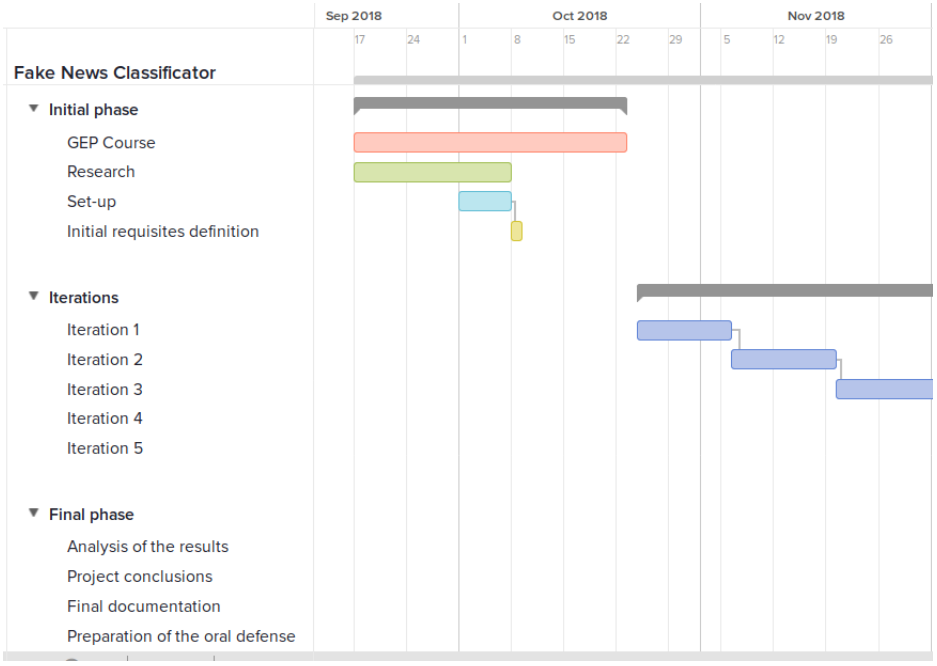
During the second week of October, after finishing the GEP course. The constant work of the subject during many days, added to an overload of work in external matters to the project, made the dedicated effort decrease during the first week and a half of implementation. Therefore, it can be considered that the first iteration was not finished and that the implementation started on October 23rd. In principle, we wanted to reduce the number of iterations from five to four, but given the lack of knowledge in the treatment of natural language, this has not been the case. Needing more time to learn than to implement has made the initial tasks slower to get the project structure firm. This fact has resulted in not being able to reach all the objectives on December 17th, otherwise, it is planned to finish on December 26th.

10.3.2 Change in the serialization of some tasks

All and delay the implementation process by finally focusing the project on a study rather than the realization of an information system. The planned task to analyze the results of the project is finally being carried out at the same time as the project is being developed, as this allows certain decisions to be made. This fact means that the delay in the completion of the implementation of the requirements does not affect the planning of the remaining tasks, since once the system is implemented, the results will already be analyzed and only the final documentation will remain.

10.3.3 Final schedule

Finally, with the explained alterations and how were face up the final perfom schedule can be observed below:



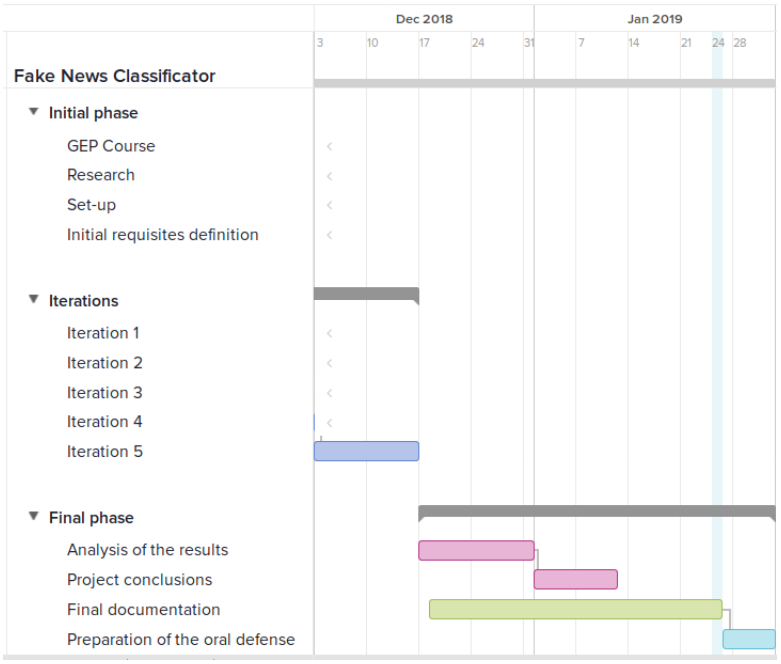


Figure 21: Final project schedule

11 Budget

On this section, all the possible costs that would appear, from being a professional project, will be explained in the following project budget planning.

11.1 Budget grouping

In this part, all the indirect and direct costs are included, take into account contingencies from the total costs. About indirect costs, hardware, software and unexpected expenses will be evaluated. On the other hand, all costs related to human activities will be included in direct costs.

11.1.1 Hardware budget

All the materials, used during the project realisation, will be included on the Hardware budget. In this case, the only element used is a Samsung laptop. The price of the laptop is estimated at 1100 euros, currently has five years, but its useful life is not much longer. Therefore, the estimate of the amortisation is five years and the duration of the project, half a year.

Product	Price	Units	Useful life	Amortization
Samsung	1100€	1	5 years	110€
Total				**110€**

Table 19: Hardware resources costs

11.1.2 Software budget

The software programs, used during the project, will be included in this budget. By the way, the total cost of the software resources, with the use of open-source tools, is zero.

Product	Price	Units	Useful life	Amortization
VSCode	0€	1	-	0€
Github	0€	1	-	0€
Latex	0€	1	-	0€
Google programs	0€	1	-	0€

Trello	0€	1	-	0€
Python libraries	0€	1	-	0€
Microsoft Translator API	0€	1	-	0€
Jupyter Notebook	0€	1	-	0€
Total				**0€**

Table 20: Software resources costs

11.1.3 Human resources budget

Although the project will be developed by one person, and to adapt the plan to reality, the author will have different roles during its realisation. In front of the list of tasks to perform, the number of hours will be estimated in function to the according to the role. The estimation of the total amount is about 536 hours, and the breakdown is shown below:

		Dedication (h)		
Activity	**Total hours**	**Project Manager**	**Analyst**	**Software Developer**
GEP Course	90	70	20	-
Research	40	10	30	10
Set-up	10	-	-	10
Requisites definition	6	4	2	-
Iteration 1	45	5	10	30
Iteration 2	45	5	10	30
Iteration 3	45	5	10	30
Iteration 4	45	5	10	30
Iteration 5	45	5	10	30
Result anlysis	60	10	20	30
Result summary	105	40	40	25
Total	536	156	152	225

Table 21: Human resources task estimation by hours

With the total number of hours that each role will work, its cost is estimated according to the salary defined for each position. The total of this cost will be the human resources budget.

Role	Total hours	Price/hour	Total cost
Project manager	159	50 €	7950 €
Analyst	152	45 €	6840 €
Developer	225	30 €	6750 €
Total	536	-	21.540 €

Table 22: Human resources costs by task estimation

11.1.4 Unexpected costs

A part of the estimated human cost is taken into account to add extra hours. Then, in the case of had done bad planning on hours estimation, this process will allow overcoming unexpected costs that may arise in the different tasks.

Role	Total hours	Price/hour	Total cost
Project manager	15	50 €	750 €
Analyst	15	45 €	675 €
Developer	15	30 €	450 €
Total	45	-	1875 €

Table 23: Unexpected costs estimation

11.1.5 Other general costs

In addition to hardware and software as indirect costs, for the work realisation, the following resources are to be taken into account. Prints, to be able to present it to the jury, and light to be able to use the hardware explained in the previous section.

Resource	Price
Prints	25 €
Light	50 €
Total	75 €

Table 24: Other general costs estimation

11.2 Total budget

Finally, the total of all cost subsets in the project are calculated. In addition, an estimated 5% of contingencies, to be able to face unforeseen expenses during the project, are included in the total.

Concept	**Extimated costs**
Hardware budget	110 €
Software budget	0 €
Human resources budget	21.540 €
Undexpected costs	1.850€
Other general costs	75 €
Subotal	**23.575 €**
Contingency (5%)	1.178,75 €
Total	**24.753,75 €**

Table 25: Total amount of project budget

12 Sostenibility

To talk about the sustainability of this project, the impact of it will be analysed guided by the three dimensions: environmental, economic and social. The analysis is done following the rules to make a sustainability matrix and the results are shown below:

	PPP	**Useful life**	**Risks**
Enviormental	5/10	20/20	-5
Economic	8/10	20/20	-10
Social	10/10	20/20	-5
Sustinability range	22/30	60/60	-20/-60
			62

Table 26: Sustainability matrix of the project

12.1 Enviormental dimension

Addressing the environmental dimension, the realization of the project will negatively impact the dimension. On the one hand, this project requires hours of electricity and objects, like the laptop, that contaminated during their creation. On the other hand, this project does not work for improving the environmental system of the world.

Is for these reasons that it does not have a good impact on this dimension. Althought there not exists any risk to consider in this dimension and the project length is short to do a big impact.

Talking about the solutions to minimize the impact would be using low-energy on electronic devices. What the author can do, to achieve that, is working on its home because is produced by clean energy.

12.2 Economic dimension

In the economic field of the project, most of the budget is going towards human resources. By using many open sources or free software resources, and not being a project that needs other more expensive Hardware devices, the budget derives from a human resource expense that goes according to the total hours of the project.

With regard to costs as an investigation, it poses a risk for uncertainty in the results. In the case of obtaining positive results, the cost of the project can be amortized could give way to an improvement in the quality of the journalism and to be able to eliminate the business that currently exists in the world of the fake news.

As there are no projects outside of research that develop this kind of purpose, it is not possible to compare that this project contributes economically more than others.

12.3 Social dimension

Finally, in the social sphere, I believe that this type of research can improve my awareness of what is happening in the information world today. and have more knowledge when analyzing the news that appears in the day to day.

Besides, it can not only improve my awareness if not that of other people, and make this project can be a tool to improve the information on society.

The false news is on the agenda and interest both economically and socially, then it is necessary any tool that can fight them.

13 Conclusions

13.1 Acquired knowledge

One of my main objectives for this project was to research and to implement about a related topic of the Computer Science but also merging it with my Software Development concepts. Two areas where I am very involved and where I also wanted to show that it is interesting to have two different points to the same problem. These two points were, in one hand, researching how is the best way to predict fake news from my knowledge and, in the other hand, how this project can be usable or prepared in order to scale it and improve it easier.

In summary, it was interesting to learn about Natural Language Processing from the base acquired during the degree. This was very unfamiliar for me and I have always been interested to know more about this area. Also, it was engaging to discover that there are a lot of different methods to understand, model and classify a group of text where their applications are infinite.

Other important thing is I have been able to apply in a real case some supervised classifiers discovered during the degree and to improve the knowledge of them. Also, I have extended my experience with the programming language called Python and the libraries for solving this type of problem as sklearn.

Finally, I could implement my first API with Flask in order to use the best classifier that I was able to find. Also, to know how to get website content with the BeatifulSoup tool.

13.2 Project results

In this project, some interesting differences on the worked sample of fake and real news could be observed. In most of the performed experiments was not possible to get good results, but in other cases some results have been revealing for a future work of improvement. It can be said that the ground has been prepared in order to find the solution to a complex problem.

The first problem found was the limited consensus about what are the fake news and how to detect them manually. This situation implied to start from a broad approach on research and not knowing or expecting that determinate certain strategies were going to work better.

About the fake news, we discovered that most of them do not come from newspaper articles. Moreover, their main feature is that all the fake news try to move the people who read the information, which it becomes to a difficult fact to extract automatically.

Analysing the performed experiments different conclusions can be drawn from the two research paths taken.

Regarding the style analysis, it was difficult to extract the information from the style of the article in order to group by their type. Low correlations were observed and as a consequence the classifiers did not return good results at all. Then, it can be concluded that the project was not able to differentiate the articles by the features extracted from the style of the articles.

However, there were some differences about the methods performed on the style experiments and it is true that a dimensional reduction of the worked data given worse results that the performed on the supervised classifier with all the dimensionality. Fact that explains that there was a lack of information using LDA and PCA.

In the experiments based on the content, it has observed very different results depending on the method. The first experiment, which was trying to classify the content by the similarity words, showed good results in order to approach the objective of making a good classification.

On the second experiment about classifying by their topic distribution is the one that showed worst results. The main reason is the lack of data, specially because each execution in the topic modelling process gives very different results and, as a consequence, it was impossible to classify.

Finally, it could have been interesting to perform a study about the improvement level of the web service classifier while it was retrained with the new data. However, the limited time of the project makes this not possible. Although it is viable to use this tool in order to consult articles, and even though it does not give perfect predictions, the classifier can give you an orientation automatically.

14 Future work

From the project results, some aspects from the worked processes can take an improvement in different ways for the future. In this section, how to approach the objectives not achieved and the new ones will be explained in order to open new paths of work.

The principal path of work thought is the one that was not possible to do because of the lack of time. This task is about to re-train the implemented web service in order to see the classifier behaviour when articles that had a bad prediction are introduced on it. And finally, analyses if expanding the dataset is a good solution to improve the accuracy of the predictions.

Another case to take in mind is raise about how important is the style classification for following this path of research. Not only because the style can be changed very fast in all type of documents, depending on what media companies see that works better for the reader, but it is also more about the extracted features. Low correlations with the label fake were observed, so although it is possible to obtain better results from a more accurate classifier maybe it is better to know about new style features to extract to see its behaviour.

And finally, about the Latent Dirichlet Allocation experiment, the one that worked worst with our data, it could be interesting increase a lot the dataset to confirm if the problem was the size or if the problem it is that is not possible to extract topics of this domain problem.

It should be noted that this project has superficially used different and very different methods to check how it behaved in each case. It can be considered the beginning of a very long way in case you want to automate its prediction because the complexity of the problem is sufficiently high.

References

[1] Maldita.es - periodismo para que no te la cuelen. https://maldita.es/. Accessed: 2018-10-24.

[2] Maldita.es dejemos de hablar de 'fake news' y de 'noticias falsas'. https://maldita.es/maldito-bulo/dejemos-de-hablar-de-fake-news-y-de-noticias-falsas/. Accessed: 2018-09-20.

[3] Periodismo, tecnología y dato. https://maldita.es/. Accessed: 2019-01-24.

[4] Study.com what is propaganda: definition, techniques, types and examples. https://study.com/academy/lesson/what-is-propaganda-definition-techniques-types-examples.html. Accessed: 2018-09-20.

[5] Wikimedia Foundation news satire. https://en.wikipedia.org/wiki/News_satire. Accessed: 2018-09-20.

[6] Garci´a Marc Amoro´s and E´vole Jordi. *Fake news: la verdad de las noticias falsas*. Plataforma, 2018.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. http://dl.acm.org/citation.cfm?id=944919.944937, March 2003. Accessed: 2019-01-20.

[8] Foro Europa Ciudadana. El debate de como combatir las fake news en las redes sociales llega al parlamento europeo. https://www.europaciudadana.org/el-debate-sobre-como-combatir-noticias-falsas-en-las-redes-sociales-llega-al-p Apr 2018. Accessed: 2018-09-20.

[9] Adam Geitgey. Natural language processing is fun! https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e, Jul 2018. Accessed: 2019-01-20.

[10] Charles J Geyer. Principal components theory notes. http://www.stat.umn.edu/geyer/5601/notes/spect.pdf, Aug 2007. Accessed: 2018-09-20.

[11] Alex Hern. Youtube to crack down on fake news, backing 'authoritative' sources. https://www.theguardian.com/technology/2018/jul/09/youtube-fake-news-changes, Jul 2018. Accessed: 2018-09-20.

[12] Benjamin D. Horne. Medium fake news starts with the title. https://medium.com/@benjamindhorne314/fake-news-starts-with-the-title-ad7b63bf79c0. Accessed: 2018-09-20.

[13] Reyson University Library. Research guides: Fake news: Identifying fake news. http://learn.library.ryerson.ca/fakenews/identify, Sep 2018. Accessed: 2018-09-20.

[14] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2017.

[15] AYLIEN Noel Bambrick. Analytics big data data mining and data science. https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html, Jun 2016. Accessed: 2019-01-20.

[16] PolitiFact. Fact-checking u.s. politics. https://ifcncodeofprinciples.poynter.org/, Dec 2018. Accessed: 2018-10-24.

[17] Slick. Commit to transparency - sign up for the international fact-checking network's code of principles. https://ifcncodeofprinciples.poynter.org/. Accessed: 2018-09-20.

[18] Max Welling. Principal components theory notes. http://www.cs.huji.ac.il/~csip/Fisher-LDA.pdf, Oct 2012. Accessed: 2018-09-20.

[19] Xinyi Zhou and Reza Zafarani. Fake news: A survey of research, detection methods, and opportunities. *CoRR*, abs/1812.00315, 2018.